A Methodology of Dataset Generation for Secondary Use of Health Care Big Data(Abstract_要旨)

AUTHOR(S):

Iwao, Tomohide

様式VI

博士学位論文調査報告書

論 文 題 目

A Methodology of Dataset Generation for Secondary Use of Health Care Big Data

（保健医療ビッグデータの二次利用におけるデータセット生成に関する方法論）

申請者氏名　　岩尾 友秀

最 終 学 歴　　平成・令和　17 年　3 月

奈良先端科学技術大学院大学

情報科学研究科 情報システム学　専攻修士課程　修了

平成・令和　2 年　3 月

京都大学大学院情報学研究科　　社会情報学　専攻博士後期課程

研究指導認定見込

学 識 確 認　　令和　　年　　月　　日（論文博士のみ）

論文調査委員　京都大学大学院情報学研究科

（調査委員長）　教　授　黒田知宏

論文調査委員　京都大学大学院情報学研究科

教　授　守屋和幸

論文調査委員　京都大学大学院情報学研究科

教　授　吉川正俊

（ 続紙 1 ）

京都大学	博士（　　情報　　学）	氏名	岩尾 友秀
論文題目	A Methodology of Dataset Generation for Secondary Use of Health Care Big Data （保健医療ビッグデータの二次利用におけるデータセット生成に関する方法論）		

（論文内容の要旨）

保健医療データは，電子カルテデータやレセプトデータなどに代表される日常の診療や調査に伴い自動的に蓄積されるデータである．現在ではほとんどの医療機関等がデータを電子化した状態でデータベースに蓄積しており，データ量に関してはビッグデータといえるケースも増えている．このような背景から，近年では電子化された保健医療データに対して日々進歩する工学領域の知見が援用しやすいことも手伝い，疾患の予防や治療法に関するエビデンスを構築することで人々の健康増進に貢献すべく，疫学研究への二次利用が期待されている．しかしながら，現状では保健医療データベースが活発に利活用されているとは言い難い状況である．

主な理由としては，国などのデータ提供者がこれらデータの二次利用に対して厳しいセキュリティポリシーや法規制を設けている場合や，データ自体の信頼性が低く二次利用に適していないというデータ分析以前の根本的な課題が挙げられる．一方で，データ分析の段階においても，一般に保健医療データベースは二次利用を想定して構築されたものではないという性質故，主に以下に述べる二つの課題が存在する．

一つ目は，疫学分析に必要となる属性がかならずしも保健医療データベースに備えられているとは限らないというデータコンテンツに関する課題である．

二つ目は，保健医療データは正規化された状態でデータベースに格納されることが一般的であるため，患者に関する情報が多くのリレーションに分散していること加えて，各リレーションにおいては同一患者に関するレコードが複数存在する場合が多くなるというデータ構造に関する課題である．

疫学分析では，研究者が提案した仮説を検証するために統計解析を実施するケースが一般的であることから，患者単位に沿って分析に必要となる属性を備えた患者単位のデータセットを作成することが望まれる．このため，研究者はデータコンテンツとデータ構造に関する課題を同時に解決しつつ患者単位のデータセットを作成する必要がある．データセットを作成するためのデータハンドリングには相当のプログラミングスキルが求められる．このような背景から，リサーチクエスチョンを持っている臨床家が速やかにデータベースを用いた疫学研究を実施することは困難であるというのが我が国も含めた世界的な共通課題である．

本研究は，プログラミングなどのデータハンドリングスキルに乏しい研究者であっても，保健医療データベースから患者単位のデータセットを自力で作成できる方法論を確立することを目的とし，保健医療データベースのひとつとして知られるレセプトデータベースを題材として研究を実施した．

本研究では，リサーチクエスチョンをひとつの命題としてとらえることで，患者ごとのデータセット作成にリレーショナルモデル理論を用いることができる点に着目した．本手法では，リサーチクエスチョンを複数の命題に分割することで単純化し，それぞれの属性を計算する過程と，リサーチクエスチョンによって定義された属性間の独立した拘束条件を真理関数により解決する過程に分離する．また，属性を計算する過程において，リレーショナルモデル理論における関係の閉包性に着目することで，それぞれの属性間に時系列などの任意の属性に関連する拘束条件が存在するケースにおいても，それぞれの属性を簡易に演算可能となるリレーショナル論理式を提案した．物理的にカバーできない属性に関しては，リレーショナル論理式に最適化したデータウェアハウスを構築することで補い，これらリレーショナル論理式とデータウェアハウスを協業させることでデータセット作成過程を簡易化することを実現している．

提案手法を用いて横断研究２件，後ろ向きコホート研究３件の計５件の疫学研究を実施したところ，臨床家が実施した２件の疫学研究に関しては，臨床家自身が自力でデータハンドリングを実施することができた．これら５件の疫学研究は，データベースを用いた疫学研究のほとんどを占めるとされる横断研究とコホート研究を含んでいる．そのため，本手法に関する有用性に関しては一定程度検証できたと考えられる．また本手法は，患者IDが含まれているデータベースであれば適用可能であるため，本研究で用いたレセプトデータベースのみならず，電子カルテや介護保険データをはじめとした多くの保健医療データベースに応用することが可能である．

本研究で確立したデータセット生成に関する方法論を様々なバックグランドを持った利用者に流布していくことで，保健医療データベースを用いた研究を実施できる人々が増え，多くの有用な研究成果が生み出されることが期待できる．また本手法は，データセット生成に汎用性の高いリレーショナルモデル理論を用いているため，大容量のデータを扱うことに適しており，ビッグデータを対象とするプラットフォームの開発基盤等に応用可能である．

（続紙 2 ）

（論文審査の結果の要旨）

本研究は、臨床研究を志す臨床家（Physician Scientist）が、NDB（レセプト情報・特定健診等データベース）などの大規模医療データを用いて疫学研究を行う際に、容易に基本的なデータセットを作成できるようにする方法論を考案し、これを実装し、具体的疫学研究を通じてその効果を実証したものである。

本研究の着眼は、疫学研究のリサーチクエスチョンを、リレーショナルモデル理論における一つの「命題」と捉えて記述することにある。疫学的に要求される様々な拘束条件等を真理関数や論理式として記述することで、データセットの容易な作成を可能にした。加えて、複数の疫学研究を分析し、NDB等の保健医療データベースに一般的な疫学分析で利用されるデータを追加しつつ、上記のリレーショナル論理式に最適化して、疫学研究に最適化したデータウェアハウスを構築した。これにより、多くの疫学研究が必要とする分析をデータベース演算のみを用いて実現することが可能となり、大規模疫学研究に必要な計算機資源の要求を縮減することにも貢献している。最後に、本研究では、5件の疫学研究の実施を通じて、提案手法を適用することにより、Physician Scientistが自ら研究を実施できることを確認した。

本研究の成果は、医療データサイエンスの裾野を拡げることに大きな貢献をするものであり、社会情報学分野の研究成果として高く評価することができる。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和 2 年 2 月 13 日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。

また、本論文のインターネットでの全文公表ついても支障がないことを確認した。

要旨公開可能日： 令和 2 年 3 月 25 日以降